DOCUMENT RESUME

ED 442 850 TM 031 270

AUTHOR Nokelainen, Petri; Ruohotie, Pekka; Tirri, Henry
TITLE Professional Growth Determinants--Comparing Bayesian and Linear Approaches to Classification.
PUB DATE 1999-04-19
NOTE 30p.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Classification; Computer Software; *Employment; Foreign Countries; Longitudinal Studies; *Professional Development
IDENTIFIERS Finland; *Linear Models

ABSTRACT

Bayesian and classical approaches to classification of vocational data were compared using an educational data set from a longitudinal study of professional growth and development in organizations (P. Ruohotie et al., 1994). Data were from 2,430 workers in companies in Finland who completed a questionnaire with behavior and background statements. The main purpose of this study was to look for new possibilities in analyzing multiform vocational data starting from the level at which traditional linear methods become too complex to apply. After describing the data and the theory of professional growth, the paper discusses linear discrimination and its use in the social sciences. Bayesian modeling with the BAYDA software package is described. It is concluded that linear and nonlinear methods support each other depending on the subject of the study. The Bayesian approach, in the form of the BAYDA program, is still under rapid development, but it appears to provide a valuable tool for analysis. (Contains 10 figures, 21 tables, and 12 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Professional Growth Determinants - Comparing Bayesian and Linear Approaches to Classification

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Petri Nokelainen and Pekka Ruohotie
University of Tampere

Henry Tirri
University of Helsinki

P. Nokelainen

ED 442 850

In this paper we apply Bayesian modeling into vocational research with the assistance of the previously described BAYDA[1] -software. We proceed by first describing the data used, and the theory of Professional Growth (Ruohotie,1996) which is applied into practice. After setting up this framework we briefly discuss linear discrimination and its usage I n the field of social sciences.. Due to the fact that the previous chapter has taken an exhaustive look into Bayesian classification in the final phase we can compare linear and non-linear methods.

The main motivation of our work here was to apply both the Bayesian and classical approach to classification for vocational data—the purpose was to show a real life example for vocational researchers which illustrates the differences between the two approaches. The point was to look for new possibilities in analyzing multiform vocational data starting from the level where traditional linear methods leaves us or becomes too complex to apply. The idea of this chapter is not to force the vocational researchers to choose one approach over the other. We are more interested in introducing the Bayesian approach to the modeling practices in vocational education, judging the usefulness of which is left for the practititoners.

Data

On this study we used an educational data set from a longitunal study of Professional Growth and Development in Organizations (Ruohotie et al., 1994). This study provides information about ongoing learning and self-development by employees aiming to prevent skill obsolescence. The basic source of information was the Growth Needs Project (Ruohotie, 1995) which cathered knowledge about professional updating and the problems and prerequisites of continual growth in various work communities. As this article concentrates mainly on statistical matters, more background information considering problem field of professional growth terminology and research can be obtained from previous publications (see Ruohotie et al., 1994, 1995, 1996; Beairisto, 1996)

TM031270

In macro level problems of professional updating and prerequisities of continual growth can be expressed according to Growth Needs Project in terms of factors within the individual, the job, the work place and society. When we give those subjects a form of a question, the list could look something like this:

[1] The BAYDA software is available from http://www.cs.Helsinki.FI/research/cosco/.

BEST COPY AVAILABLE

- What kinds of opportunities, supports and rewards do different work communities have for the continual development of professional qualifications?
- How can the conditions which motivate and support growth in a work community be improved?
- What connection is there between a supervisor's leadership abilities and the growth conditions prevailing in a work community?
- Which factors motivate people in different types of organizations to be goal-oriented and to develop their own expertise on a continual basis?
- What are the real effects of developmental activities in the short and long term?

The data set used in this study was gathered for the The Growth Needs Project through period of time from 1991 to 1996. (Table 1.)

Table 1

The description of the data set used.

Data set	Size	Number of Variables	Number of Classes (groups)	Total number of Variables
Professional Growth	2430	56	5 (2,2,3,4,5)	61

The purpose of the research project was to compose a valid view to the study field of professional updating through several sub-studies e.g. Lahti-Kotilainen, (1992) and Kautto-Koivula (1993). Results of the study were immediately applied in practice - in both business and educational settings (Ruohotie, 1996).

Basic model of Professional Updating was originally operationalized into set of 70 variables, but since the instrument has gone through a lot of evalution sequences we have now here 56 variables which are common to all evaluation instruments. A more detailed description of the framework and research conducted in the project is discussed in (Ruohotie, 1994).

The evaluation instrument consists of 56 behaviour and 5 background statements. Workers who filled the forms represent three Finnish companies which area of interest lies on food, vehicles and cleaning service. Instrument has Likert scale from 1 to 5, except for bCackground variables. The methods and results of this study are reported in detail in numerous articles of professor Ruohotie and his colleagues (1992, 1994, 1995, 1996).

The data used in this study was gathered from workers of two middle class companies and one larger size company in Finland (N=2430). Hence one of the most interesting classifier was the company (comp). To prevent too straightforward interpretations we tested also four other variables discriminating level of education (educat), gender (gender), age (age) and job profile (title). A description of the background variables can be found in Table 2.

Variables that mirror Triggering factors in Professional Development Process (Ruohotie 1996, 25) are divided in three groups (see Table 3). First group (Organization) includes statements comprising mission, organizing, hierarchy and working conditions of

an organization. This group covers also statements considering superiors ability to share responsibility, ideas and rewards. Second group of statements (Work Role) describes employees position as a group member and his or her relationship towards other colleagues. Work role covers also employees conception of his or her work. These statements try to clarify if employee finds his or her task challenging and rewarding. In addition professional appreciation among colleagues is measured with statements like "I feel that my work is appreciated." Third group (Person) is pure self-esteem indicator. The idea is to find out either employee is eagerly working towards his or her own goals or not.

Table 2

The description of background variables.

No.	Variable Name	Variable Description
1.	Gender	1 = Female
		2 = Male
2.	Age	1 = < 25 years
		2 = ≥ 25 years
3.	Educat	1 = Elementary School
		2 = Secondary School Graduate
		3 = Vocational School
		4 = College / Training Center
		5 = Academic Degree
4.	Title	1 = Worker
		2 = Official (no subjects)
		3 = Inferior Manager
		4 = Manager
5.	Comp	1 = Company A, middle size (N=1000)
		2 = Company B, middle size (N=919)
		3 = Company C, middle size (N=511)

Table 3

The description of triggering variables.

No.	Variable Group	Variable Description
		(1 = Strongly disagree, 2 = Disagree, 3 = No opinion ,4 = Agree, 5 = Fully agree)
6. – 25.	Organization	Statements comprising organization and acting superiors.
26. – 47.	Work Role	Statements covering work group and job.
48. – 61.	Person	Statements comprising employees person.

Theory of Linear Discrimination

Next we set out to examine linear discriminant analysis as a tool for classifying cases into different groups with a better than chance accuracy. We also study variable detection (eject/reject) by comparing LD (Linear Discrimination) with SPSS 7.0 for Windows and NLD (Non-linear Discrimination) with BAYDA (Silander and Tirri, 1998).

Discriminant analysis is close to both ANOVA and MANOVA. In precedent case one can ask wether or not two or more groups are significantly different from each other with respect to the mean of a particular variable. In posterior case we ask wether group membership is associated with reliable mean differencies on a combination of dependent variables.

The linear combination of variables (discriminant function) is similar to the right side of a multiple regression equation because it sums the products of variables multiplied by coefficients. Then procedure estimates the coefficients and the resulting function can be used to classify new cases. Next we will shortly describe the four main applications of linear discriminant function analysis.

Variable Selection

The most common application of discriminant function analysis is to let selection method to determine the subset of variables which constitutes relevant model. Entry or removal decision can be made with forward, backward or stepwise method.

Forward Selection

Forward selection first removes all variables from the model and after that starts enter them one by one. The first variable entered at step one is the one with the strongest correlation with the dependant (classification) variable. At each subsequent step the variable with the strongest partial correlation enters the model. The hypothesis that the coefficient of the entered variable is 0 is tested using its F statistic. Stepping stops when an established criterion for the F no longer holds.

For each candidate predictor variable, F statistic is computed that measures the change in Wilks' Lambda when variable is added to the list of accepted variables. The variable with the largest F enters the list. We just announced that "*stepping stops when an established criterion for the F no longer holds*". Now it is time to check what exactly is that established criterion.

The F value for the change in Wilks' Lambda when a variable is added to a model that contains p independent variables is

$$F_{change} = \left(\frac{n - g - p}{g - 1} \right) \left(\frac{(1 - \lambda_{p+1} / \lambda_{p})}{\lambda_{p+1} / \lambda_{p}} \right)$$

where n is the total number of cases, g is the number of groups, λ_p is Wilks' Lambda before adding variable, and λ_{p+1} is Wilks' Lambda after inclusion (SPSS Inc., 1997).

Backward Selection

Backward method is opposite to forward selection. The idea is to start from situation where all variables are in the model and after that, step by step, removing least useful predictor, end up in a situation where only the strongest predictors exist.

Stepwise Selection

Stepwise selection is identical to forward selection except for one point, at each step already entered variables are tested for removal. This makes sense because f.e. entry of third variable can diminish the importance of an already entered variable. We use in this experiment stepwise selection for SPSS and forward selection for BAYDA.

Even though mathematically MANOVA and linear discrimination are the same, classification is a major extension of linear discrimination over MANOVA. Point is to find out how well we can predict to which group a particular case belongs.

Importance of variable selection becomes more important when we discuss about problem of overfitting. In that case one has to be aware not to include too many variables in the model. The problem seems to be the fact that such a model will not predict correctly when applied to a new sample.

We might draw a conclusion about variable selection that selection methods are useful tools to find which are the most important variables describing the phenomenon we are studying. These tools also help us to drop down the number of variables to help us avoid the problem of overfitting. As we analyze the results of variable selection we might consider the method that comes along with least number of variables as the best predictor. This all comes down to fact that it is a lot easier to report common factors of five than twenty variables.

Two-Group Discriminant Function

Primary goal for many researchers that use discriminant analysis is to find discriminant function to predict group membership. The most important issues we can ask are: *Can group membership be predicted reliably from the set of predictors? What is the number of significant discriminant functions? What are the dimensions of Discrimination?*

Criteria for evaluating overall statistical reliability are based on multivariate tests e.g. Wilks' Lambda, Hotelling's trace criterion and Pillai's criterion. Here we concentrate on describing major features of Wilks' Lambda (see equation below). Wilks' Lambda is a likelihood ratio statistic that tests the likelihood of the data under the assumption of equal population mean vectors for all groups against the likelihood under the assumption that population mean vectors are identical to those of the sample mean vectors for the different groups. When testing equality of groups centroids it varies between 0 and 1. Small values indicate that the group means differ.

$$\Lambda = \frac{|S_{error}|}{|S_{effect} + S_{error}|}$$

That makes Wilks' Lambda as the pooled ratio of effect variance to error variance to effect variance plus error variance (SPSS Inc, 1997; Tabacnhick et al., 1996). On comparison, Pillai's criterion is simply the pooled effect variances. When separation of

groups is distributed over dimensions, Pillai's criterion is more adequate. Most research reports use Wilks' Lambda unless there is reason to use Pillai's criterion.

For two groups one discriminant function (often called Fischer discriminant function after R.A.Fischer) is computed, for three groups two discriminant functions are possible. We will discuss multiple groups discriminant analysis later on this chapter.

As stated earlier, computationally linear discriminant function analysis is analogous to MANOVA. For two variables, discriminant function is an equation of a plane where we fit following linear equation between two groups:

$$D = d_1z_1 + d_2z_2 + \ldots + d_mz_m$$

Discriminant function score (D) is found by multiplying the standardized score on each predictor (z) by its standardized discriminant function coefficient (d) and the adding the products for all predictors. (Tabachnick et al., 1996)

Multiple Groups Discriminant Functions

When analyzing more than two groups, one has to focus on canonical variables. The first canonical variable is the linear combination of the variables that maximizes the differences between the means of the *n* groups in one dimension. The second canonical variable represents the maximum dispersion of the means in a direction orthogonal to the first direction.

Interpreting canonical functions is somewhat similar to factor analysis, comparing canonical variables as factors that discriminate optimally among the group centroids relative to the dispersion within the groups. There are two main paths to examine canonical functions, first one can compare (e.g. in a table) the second canonical variable against the first. This provides an easy way to display group differencies. Second path is to look factor structure for which variables define best a particular discriminant function. The factor structure coefficients are the correlations between the variables in the model and the canonical functions.

On Table 4 we present example of using canonical variables to display of group differencies. Next figure (Figure 4) is a scatterplot of first and second canonical variables displaying Organizational Triggers (see Table 3) in three different companies (see Table 2).

Table 4

Canonical Discriminant Functions at Group Centroids.

	Canonical Function	
Company	**1**	**2**
Company A	,744	,948
Company B	,558	-1,095
Company C	-2,460	,115

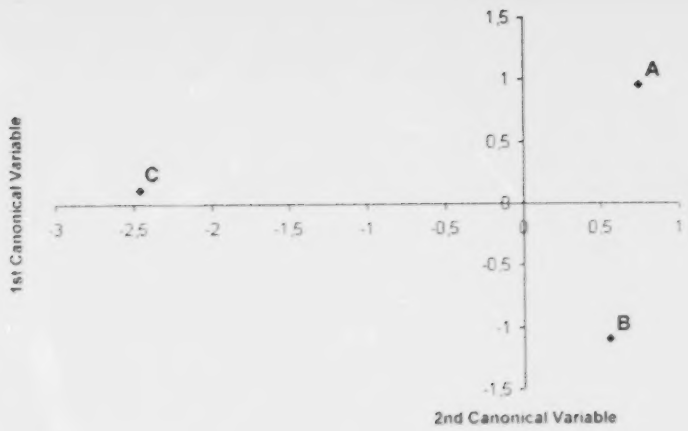


Figure 1

Canonical Variable Plot of Company – Organizational Triggering Variables.

The plot displays information about differencies in three organizations. Second canonical function discriminates clearly between Company C and other two companies.

Classification

After discriminant functions have been derived we can predict with classification functions to which group particular case belongs. Classification functions are not to be confused with the discriminant functions. There are as many classification functions as there are groups. Each function allows us to compute classification scores (which can be stored as new variables) for each case for each group with the following formula:

$$C = c_0 + c_1X_1 + c_2X_2 + \ldots + c_mX_m$$

A score on the classification function group (C) is found by multiplying the score on each predictor (X) by its associated classification function coefficient (c), summing over all predictors, and adding a constant c_0. (Tabachnick et al., 1996)

When considering predictive classification of cases one must ask: *What is adequacy of classification?* This means in the first place that researcher must evaluate the classification results based on his/her previous knowledge on subject. Another way to approach this problem is to compare cases correctly classified by classification procedure to those obtained by chance alone. This is done simply by making allegation that 50% of cases are correctly classified by chance alone when there are two groups and 33% with three groups. This conclusion assumes that all groups are equal by size.

Linear Approach to Classification

Background Information

As we now understand, the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of variable. That variable is in turn used to predict group membership. As the goal of discriminant function analysis is to predict group membership from a set of predictors, we can for example study if climate of organization or gender are valid predictors for Professional Growth.

Naturally at this stage we must ask question from ourselves that clarifies the usage of classification: *What answers do we obtain by applying linear discrimination to the vocational data?*

First, we gain valuable information about variables predicting effectiveness of Professional Growth Triggering variables. Second major factor is predictive information. After model validating through test measurements we can obtain same questionnaire to new organizations[2] to find out wether it has certain characteristics or not. These special characteristics can be used as indicators of professional updating.

Let us now first study the problem of developing a classification procedure, which would allow us predict the group to which a given data vector most likely belongs. When one is interested in forming a hypothetical model describing life-career's development as a professional in a view of growth, it is clear that classification based on organization is easy to find thrilling.

As we study here living organizations trying to build a solid theory of professional growth development, one way to operationalize the problem of prediction is to measure real organizations based on theory of professional growth and then validate findings by observing.

Here we will allow the classification procedures to use 20 Organizational Triggering Variables in constructing the predictive model. In practice for this type of problems discriminant analysis is preceded by dimensionality reduction procedures, e.g., factor analysis, and one would use summarized information such as the factor scores instead of the primary variables.

Knowing the difficult issues related to selecting a proper factor structure, this would, however, introduce another parameter to our study, i.e., the discriminative quality of the factor variables constructed. Although the analysis is performed at the primary variable level, all discussion is also naturally valid for discriminant analysis with factor scores.

Before we perform discriminant analysis, careful data analysis has to be performed. This means that we explore critically all cases and exclude all those that will not meet allegation of normality.

---

[2] In this case 'new organization' is Company C. We use same set of organizations twice, first measuring grouping of COMP variable and after that two grouped COMP variable again. Later on we include new organizations in the data.

Definitions

First we load the data (N=2430) in the SPSS® 7.5 for Windows. After that follows parameter selection, we choose **Statistics – Classify – Discriminant...** Following window (Figure 2) is displayed:

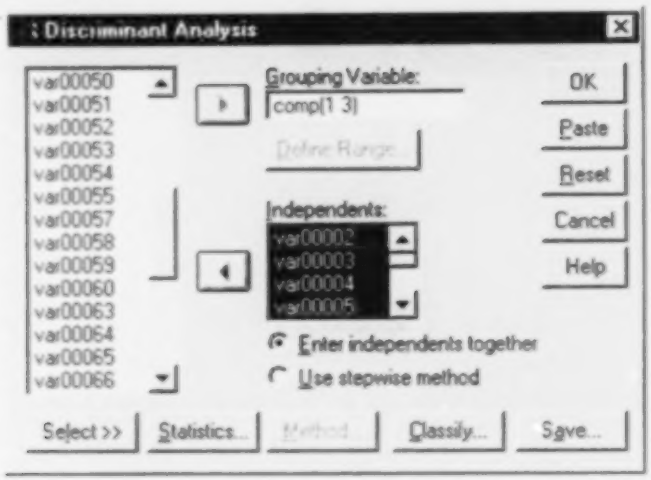


Figure 2

Discrimination Dialog Box in SPSS® 7.5 for Windows

We select following settings in **Discrimination Dialog Box**:

- Grouping Variable: COMP
  Define Range...
  Minimum: 1
  Maximum: 3
- Independents: variables 6 – 25 (Organization)

× Use stepwise method

Statistics...
- Descriptives
  - × Means
- Function coefficients
  - × Fischer's
- Matrices
  - × Within groups correlation

Method...
- Display
  - × F for pairwise distances

Classify...
- Display
  - × Summary table
  - × Leave-one-out classification
- Plots
  - × Combined-groups
  - × Separate-groups
  - × Territorial map

Save...
- × Predicted Group Membership

Tests of Equality of Group Means

Tests of equality of group means panel (Table 5) provides information regarding differencies among variables. All variables vary significantly as column six indicates.

BEST COPY AVAILABLE

The strength of the intercorrelations among the variables is important. On the other hand, if variable has low level of significance it is not necessarily reason to exclude it from analysis.

Wilks' Lambda gives information regarding differencies among groups. As we can see, column two indicates that there are no strong group differencies. Big sample size may have an effect on this, because sixth column still indicates acceptable significance levels.

Table 5

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig
VAR00002	,940	78,038	2	2427	,000
VAR00003	,966	42,955	2	2427	,000
VAR00004	,835	239,834	2	2427	,000
VAR00005	,958	53,728	2	2427	,000
VAR00006	,900	134,618	2	2427	,000
VAR00009	,988	15,226	2	2427	,000
VAR00012	,946	68,896	2	2427	,000
VAR00013	,995	6,487	2	2427	,002
VAR00014	,944	71,723	2	2427	,000
VAR00017	,819	268,228	2	2427	,000
VAR00019	,978	27,791	2	2427	,000
VAR00020	,843	225,626	2	2427	,000
VAR00021	,906	126,185	2	2427	,000
VAR00022	,861	195,096	2	2427	,000
VAR00024	,780	342,378	2	2427	,000
VAR00028	,953	60,465	2	2427	,000
VAR00030	,878	168,988	2	2427	,000
VAR00031	,981	23,209	2	2427	,000
VAR00032	,842	228,144	2	2427	,000
VAR00034	,994	7,519	2	2427	,001

Variables in the analysis

SPSS accepts by using stepwise method 19 out of 20 variables into analysis. We discussed earlier on this chapter meaning of F –value to accept or reject variable from analysis, but we still underline that F value gives changing information on different variables during selection and it is informative to follow tolerance of different variables (e.g. when entered/removed).

Eigenvalues

First Eigenvalue (canonical discriminant function) account for 66,2% of the total dispersion (Table 6). This means that it corresponds to the canonical discriminant function in the direction of the maximum spread of the group means. We may generalize that smaller the eigenvalue, less account for the total dispersion. By observing 3rd column we notice that Second canonical discriminant function has 33,8% of the spread of dispersion.

5th column on table below presents the correlation between each canonical discriminant function and the dummy set of variables defining the structure of the groups.

Table 6

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	1,620[a]	66,2	66,2	,786
2	,827[a]	33,8	100,0	,673

a. First 2 canonical discriminant functions were used in the analysis.

Wilks' Lambda

On Table 7 observed significance level (Sig.) is less than 0.0005. On the basis of that information we can reject the hypothesis that group centroids (means) are equal. When the first function is removed, Wilks' Lambda is 0.547 and *p* value is still below 0.0005. This means that it is worth keeping both functions.

Table 7

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,209	3785,448	38	,000
2	,547	1456,816	18	,000

Canonical Variables

Table 8 displays canonical variables. We can compute canonical variable score for each *case*:

score = 0.179*VAR00002* – 0.107*VAR00003* + 0.492*VAR00004*... - 1.336

The signs of the coefficients have no effect on separation. The number of canonical variables is $k - 1$ (k is the number of groups). Thus we have two discriminant functions for three groups Table 9 presents the same coefficients standardized.

Table 8

Canonical Discriminant Function Coefficients

	Function	
	1	2
VAR00002	,179	-,315
VAR00003	-,107	,157
VAR00004	,492	,041
VAR00005	-,189	,282
VAR00006	,304	,095
VAR00009	-,027	-,282
VAR00012	,064	,103
VAR00013	-,166	,037
VAR00014	,246	-,385
VAR00017	-,670	,010
VAR00020	,119	,423
VAR00021	-,021	,204
VAR00022	,342	-,019
VAR00024	,158	,515
VAR00028	,238	,014
VAR00030	-,160	,419
VAR00031	,096	-,650
VAR00032	-,772	-,105
VAR00034	,189	-,148
(Constant)	-1,134	-1,336

Table 9

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
VAR00002	,162	-,286
VAR00003	-,098	,144
VAR00004	,504	,042
VAR00005	-,205	,306
VAR00006	,298	,093
VAR00009	-,029	-,301
VAR00012	,064	,104
VAR00013	-,179	,040
VAR00014	,242	-,379
VAR00017	-,691	,010
VAR00020	,116	,413
VAR00021	-,022	,215
VAR00022	,365	-,020
VAR00024	,173	,565
VAR00028	,255	,015
VAR00030	-,178	,465
VAR00031	,109	-,742
VAR00032	-,827	-,112
VAR00034	,202	-,158

Structure Matrix

For each variable in Table 10, an asterisk marks its largest absolute correlation with canonical function. Structure Matrix can be interpreted (as discussed earlier) in the same way than factor loadings. On a Figure below we can see, that (loadings are ordered) variables *VAR00017* (*This company is willing and capable to take ideas from workers*) and *VAR00004* (*Staff has chance to improve ones work and working environment*). Second function is obviously dealing with leadership, because two stronges variables are *VAR00024* (*My superior gives me feedback about my work*) and *VAR00020* (*My superior shares response to employee*).

Notice that all variables are displayed, not just selected. Rejected varible *VAR00019* is marked with small "a".

Table 10

Structure Matrix

	Function	
	1	2
VAR00017	-,340*	,204
VAR00004	,335*	,140
VAR00032	-,327*	,132
VAR00022	,273*	,220
VAR00006	,236*	,160
VAR00002	,198*	-,032
VAR00014	,186*	-,062
VAR00009	,088*	-,013
VAR00024	,098	,568*
VAR00020	,123	,442*
VAR00030	-,135	,365*
VAR00021	,024	,353*
VAR00028	,061	,230*
VAR00005	-,039	,225*
VAR00012	,132	,186*
VAR00003	,085	,170*
VAR00019[a]	,096	,143*
VAR00031	-,041	-,141*
VAR00034	,024	,080*
VAR00013	-,016	,077*

Functions at Group Centroids

Table 11 presents canonical variable means by group. The difference among the centroids is tested for each pair of groups and plotted on Figure 3.

Table 11

Functions at Group Centroids

Company	Function	
	1	2
COMPANY A	,744	,948
COMPANY B	,558	-1,095
COMPANY C	-2,460	,115

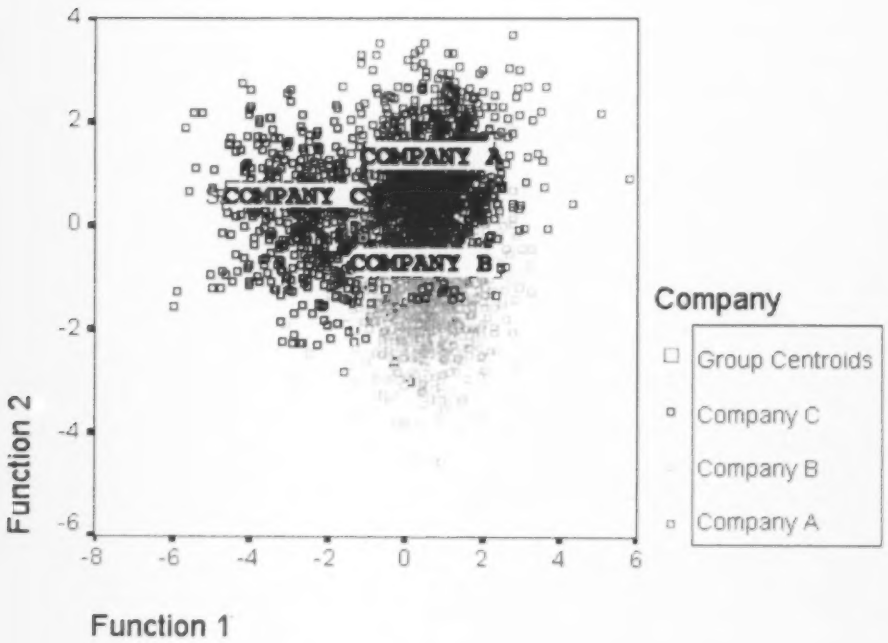


Figure 3

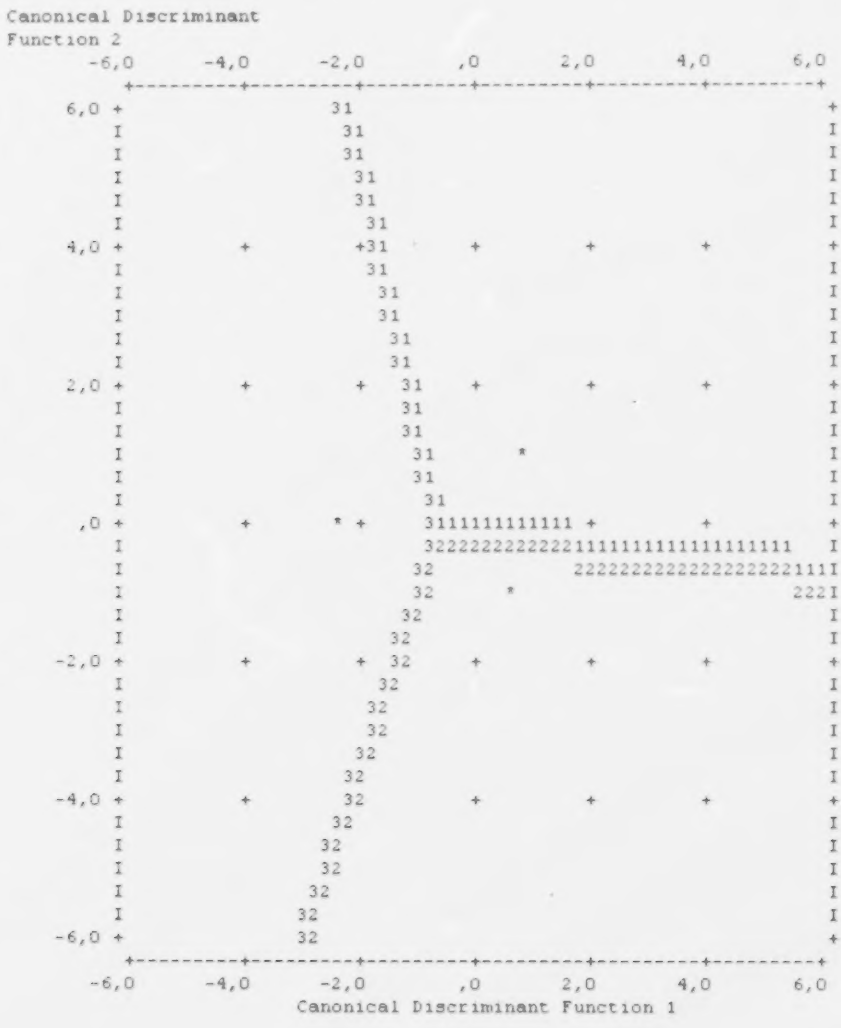
Canonical Discriminant Functions

COPY AVAILABLE

Territorial Map

Territorial Map (Figure 4) displays almost same information than canonical variable plot (Figure 3) but includes numbered boundaries marking regions into which each group is classified. For example all points to the left of number 3 are classified to group 3 (Company C).

The asterisks (*) mark group centroids.



Symbols used in territorial map

Symbol	Group	Label
1	1	COMPANY A
2	2	COMPANY B
3	3	COMPANY C
*		Indicates a group centroid

Figure 4

Territorial Map

BEST COPY AVAILABLE

Classification Results

The overall success of this 19 variable model for classifying cases into three groups is 84,4%. This is quite good result because on chance alone the result would be something like 33%. Notice that classification works best for Company C (89,2%). Company B has the highest miscalssification rate (8,7%), which is by all means still very low.

Table 12

Classification Results

			Predicted Group Membership			
		Company	COMPANY A	COMPANY B	COMPANY C	Total
Original	Count	COMPANY A	849	126	25	1000
		COMPANY B	147	747	25	919
		COMPANY C	20	35	456	511
	%	COMPANY A	84,9	12,6	2,5	100,0
		COMPANY B	16,0	81,3	2,7	100,0
		COMPANY C	3,9	6,8	89,2	100,0

a 84,4% of original grouped cases correctly classified.

Our main purpose at this stage was to show that three companies we have measured can be classified easily. We also wanted to show that Company A which was defined as professionally updating thru Triggering variables, differs from Company B. Always when object of study is a living organism we must be very cautious with change over time. If we want this measurement model to be predictive, we must 'update' it time to time with new samples of companies which present (e.g. according to theory of Organizational Triggers) neither good or bad examples of professional updating. Ideal situation is to observe few old and few new organizations in e.g. three years period. We continue linear solution later on this chapter after we perform Bayesian analysis.

Non-linear Approach to Classification

Here we perform exactly the same classification procedure than previously with linear methods. All *cursive* text is taken straight from BAYDA. We present almost identical information to that gained from linear methods. As previous chapter (Silander et al., 1998) presented walkthrough of classification in BAYDA, we may start straight from step 5 (Analyze Results).

Step 5 Analyzing Results

The results of the Bayesian predictive discriminant analysis (produced by BAYDA) are represented at three levels of details: general, groupwise and individual.

The Model for Classifying COMP

> *The task was to try a model for classifying the data items according to the class variable "COMP" using the predictor variables VAR00002, VAR00004, VAR00014, VAR00017, VAR00024, VAR00031 and VAR00032. How successfull the task was can be determined by the following report.*

As linear method chose 19 variables out of 20, BAYDA comes along with seven variables.

General Classification Accuracy

> *It can be estimated that using the selected predictor variables 74.0% of the classifications will be correct. This estimation is based on the following external leave-one-out crossvalidation procedure: Using the selected predictor variables, we built 2430 models. Each of these models were constructed using 2429 data items from the data set and each model was then used to classify the data item not used in the model's construction. Since 1798 out of 2430 models succeeded in classifying the one unseen data item correctly, one may assume that this would happen in the future as well.*
>
> *However, simply stating the classification performance of 74.0% is not meaningful as such. It has to be compared with the performance obtainable by a "default" classification procedure that always guesses the class of the data item to be the class of the majority (class "1" in this case). This simple method would yield the performance rate of 41.2%.*

BAYDA gives clear explanation how to interpret general classification results. Here we have general classification accuracy of 74.0% against dummy result (41.2%). More information on classification performance of BAYDA is available at the end of this chapter.

Classification by groups

Classification performance and its reliability by groups

> *The overall result of 74.0% is just an average performance rate. Suppose our model classifies a certain data item to belong to the class "1". Does this mean that there is 74.0% chance that this classification is right? Not necessarily, since some classifications may be correct more often than the others. In this case, while doing the crossvalidation, we predicted 957 times that data item should belong to the class "1" and 74.0% of these classifications were correct. So we (somewhat naively) estimate that if the system predicts previously unseen data item to belong to the class "1", there is 74.0% chance that this prediction is right. The reliability of this*

estimate can be rated by stating the fact that the estimate is based on classifying 957 items (39% of the sample) as members of the calss "1". Below you can find the barcharts describing the estimated correctness of different classifications. Below each estimate there is barchart indicating the percentage of the sample size used to calculate this estimate. If this estimate is based on very few predictions it is of course not very reliable.

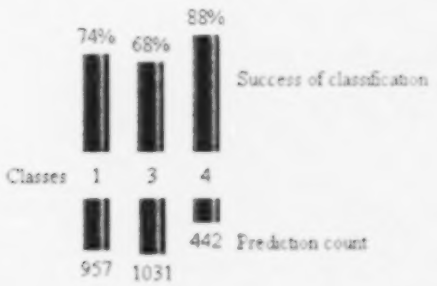


Figure 5

The estimates of the classification success and the reliability of the estimate.

Group Difficulty

Like some classifications are more reliable than the others, the data items of some classes seem to be easier to classify than the others. For example during our crossvalidation we noticed that out of 1000 data items belonging to the class "1", 708 (71%) were correctly classified. The results telling how well the data items of different classes can be predicted are represented by a Figure 6. In the same figure there are also barcharts indicating the relative class sizes.

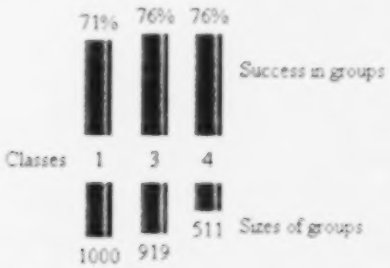


Figure 6

The group difficulty and the sizes of the groups in the sample

Individual classifications

Sometimes it is also interesting to see how well individual data items can be classified. On one hand this reveals outliers and in-doubt units, on the other hand this tells about the models we manage to construct. The results of individual classifications below were collected during the crossvalidation. For each data item the table contains the probabilities that data item belongs to different classes. In crossvalidation these probabilities were estimated using the model constructed without the data item to be classified. In each row the probability of the correct class is emphasized and the probability of the class predicted by the model is boldfaced. When these too markings do not coincide, the data item was misclassified

and its ID is marked with a red ball. For example the data item 1 has correctly been classified as belonging to the class "1", since the model has estimated that the probability of this class is highest (73%). On the other hand the data item 3 (of the class "1") has been misclassified as belonging to the class "3".

Data ID	Class probabilities (%)		
	1	3	4
1	***73***	26	0
2	***44***	42	14
3 ●	*48*	**51**	0
. . .	Follow the link to see the whole table		
2430	11	***86***	2

Figure 7

Individual classifications

Figure 7 is just a sample from table presenting all 2430 cases. Red ball points out that data vector 3 was misclassified. This is very clear and simple tool for scientists to pick up outliers, cases that are hard to classify.

Comparing Bayesian and Linear Approaches to Classification

Here we take closer look into variable selection and classification. Our purpose is to compare linear and non-linear methods as tools to produce both meaningful variable selection and classification.

Requirements for the Data

Comparing pretensions issued on data in linear and non-linear methods is easy due to fact that BAYDA has none. On the opposite, when testing data for linear methods we must be very careful. Predictor variables should be quantitative and follow normal distribution. Data should also be screened graphically with boxplots of the within-group distributions of each variable. When working with linear methods we must also take care of variances, transformations and relations among variables. We also use scatterplots to study relations among pairs of variables. Printing covariance matrix also helps us to compare between different variables across the groups.During this process we reject all variables that can not keep up with requests.

Major benefit of BAYDA over any linear statistic package is the ability to analyze almost any kind of data. There is no 'invalid' data for BAYDA. Researcher saves a lot of time and energy when he / she dont have to pick unwanted variables or cases out of the data before even planning statistical operations. We must remember at this point that if main goal is classification, linear methods are more relaxed about the data. Basic distinctions between the two methods are presented in Table 13.

Table 13

Comparing Limits to Linear and Non-linear analysis.

	Linear discriminant function analysis by SPSS® 7.5 for Windows	Non-linear classification by BAYDA
Sample size	At least 20 cases in smallest group.	At least 2 cases.
Unequal Sample Sizes	No effect. Highly unequal sample sizes are not recommended for classification.	No effect.
Missing Data	Reflects as a problem of unequal *n*	No effect.
Multivariate Normality	Normal distribution. No skewness allowed.	No effect.
Outliers	Major effect. Test for univariate and multivariate outliers for each group separately must be performed.	No effect.
Linearity	Linear relationships assumed. Violation leads to reduced power.	No effect.

Variable Selection

Tables 14 to 16 show difference between linear and nonlinear methods which carries thoughout this research; BAYDA has the ability to clarify variable selection (19 vs. 7, 10 vs.5, 10 vs. 8). By selecting entered variables efficiently one can avoid the problem of overfitting. This is very valuable feature especially for researcher who operates in a field of education.

As discussed earlier, order of entered variables is also a great source of information. On Table 13 we can notice that three out of five variables are same in both applications. Those must be good predictors for calssification.

Table 14

Selecting Organizational Triggering Variables that Discriminate Company.

Company	Linear discriminant function analysis By SPSS	Non-linear discrimination by BAYDA
Variables in the Analysis, N	20	20
Variables Selected, N	19	7
Variables, 5 first in order of appearance	24	2
	32	4
	4	14
	17	17
	31	24

Table 15

Selecting Personal Triggering Variables that Discriminate Title.

Title	Linear discriminant function analysis by SPSS	Non-linear discrimination by BAYDA
Variables in the Analysis, N	14	14
Variables Selected, N	10	5
Variables, 5 first in order of appearance	80	66
	66	69
	69	72
	74	77
	72	80

Table 16

Selecting Triggering Variables Describing Work Role that Discriminate Age (Working Experience).

Age (Working Experience)	Linear discriminant function analysis by SPSS	Non-linear discrimination by BAYDA
Variables in the Analysis, N	22	22
Variables Selected, N	10	8
Variables, 5 first in order of appearance	51	35
	35	38
	59	51
	60	52
	38	58

Classification

Classification is in its simplest form a number which tells how many percent of cases were correctly placed in their groups. This information is valuable when evaluating a model, if correct classification percent is low (e.g. below result gained by chance alone), the model applied to data is presumably inadequate. In practice this means that variables do not operationalize theoretical model correctly or the teory is false.

Tables 17 to 19 show a clear tendency when selecting variables in BAYDA; variable selection has always positive influence on general classification results. If we take closer look at Table 18 we notice that classifying result rises from 66 to 81 percents. If we also look at previous Table 14 one can see that variable selection reduced number of variables from 14 to 5.

Table 17

Classifying Company by Organizational Triggering Variables.

Company	Linear discriminant function analysis by SPSS		Non-linear discrimination by BAYDA	
No variable selection				
General classification, %	**84**		**70**	
Classification by groups, %	85	Company A	71	Company A
	81	Company B	64	Company B
	89	Company C	86	Company C
Variable selection				
General classification, %	**84**		**74**	
Classification by groups, %	85	Company A	74	Company A
	81	Company B	68	Company B
	89	Company C	88	Company C

Table 18

Classifying Title by Personal Triggering Variables.

Title	Linear discriminant function analysis by SPSS		Non-linear discrimination by BAYDA	
No variable selection				
General classification, %	**61**		**66**	
Classification by groups, %	64	Worker	94	Worker
	46	Official (no subjects)	32	Official (no subjects)
	38	Inferior Manager	12	Inferior Manager
	60	Manager	13	Manager
Variable selection				
General classification, %	**61**		**81**	
Classification by groups, %	64	Worker	87	Worker
	45	Official (no subjects)	40	Official (no subjects)
	39	Inferior Manager	19	Inferior Manager
	49	Manager	20	Manager

Table 19

Classifying Age (Working Experience) by Triggering Variables describing Work Role.

Age (Working Experience)	Linear discriminant function analysis by SPSS		Non-linear discrimination by BAYDA	
No variable selection				
General classification, %	**70**		**68**	
Classification by groups, %	74	< 25 years	51	< 25 years
	68	25 -> years	78	25 -> years
Variable selection				
General classification, %	**70**		**73**	
Classification by groups, %	56	< 25 years	74	< 25 years
	78	25 -> years	68	25 -> years

Carrying Analysis on with Linear Method

As we recall, two out of three companys (Company A and B) were involved as training sample. They 'teach' classification functions to separate between recognizable Triggering elements in different organizations. The third company (Company C, N=511) was involved because we wanted to test classifying power of two other companies. Next we do a series of tests that will tell what is situation in the third organization.

First we establish new summary variables (S_ORGAN, S_PERSON and S_WORKRO) which are formed from variables presented in Table 3. We measure three types of triggering variables with three nominal scales varying from 1 to 5 where 5 is the highest (optimal) score:

Table20

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
S_ORGAN	511	1	5	3,00	,55
S_PERSON	511	2	4	2,81	,27
S_WORKRO	511	1	5	2,98	,60
Valid N (listwise)	511				

Table 20 shows us that this organization has quite low Triggering levels varying from 2.81 to 3.00. Statistical significance can not be judged alone with those indicators. That leads us to use t-test for measuring difference with summary variables and classifying variable (generated earier in discriminant analysis).

5th column on Table 21 shows that all new variables differ significantly from grouping variable.

Table 21

Independent Samples t-test

		Levene's Test for Equality of variances		t-test for Equality of Means						
									95% Confidence Interval of the Mean	
		F	Sig.	t	df	Sig. (2 tailed)	Mean Difference	Std. Error Difference	Lower	Upper
S_ORGAN	Equal variances assumed	49.141	.000	10.967	1917	.000	.38	3.42E-02	.31	.44
	Equal variances not assumed			11.079	1870.687	.000	.38	3.39E-02	.31	.44
S_PERSON	Equal variances assumed	41.595	.000	23.306	1917	.000	.31	1.33E-02	.28	.34
	Equal variances not assumed			23.510	1888.009	.000	.31	1.32E-02	.28	.34
S_WORKRO	Equal variances assumed	12.098	.001	10.899	1917	.000	.33	3.04E-02	.27	.39
	Equal variances not assumed			10.959	1913.149	.000	.33	3.02E-02	.27	.39

Third phase is to proceed to discriminant analysis. This time we classify the data with new COMP(1, 2) variable. Our goal is to achieve classification for Company C. Figures 8 to 10 show that one can find clear trend on response to Triggering factors. New Company C is at this stage classified as "No Professional Updating", but as we want to underline, this study is experimental by its nature. Our main purpose is to show how easy it is to carry out variable classification with BAYDA. This section gives some future ideas for how to direct our endeavour to build comprehensive selection of non-linear statistic tools.

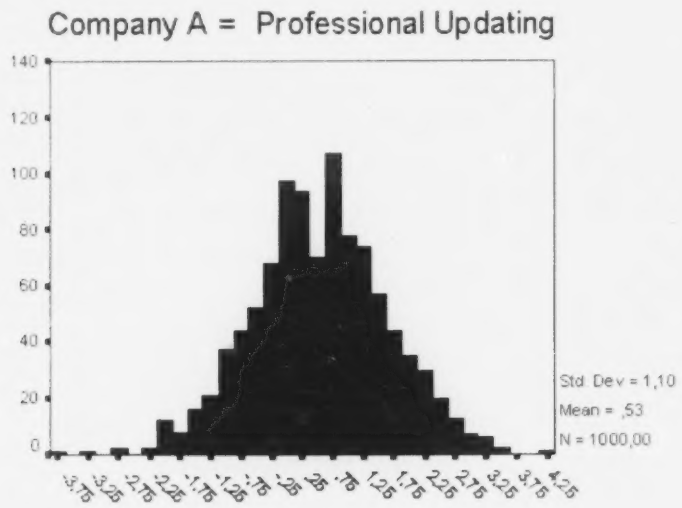


Figure 8

Canonical Discriminant Function 1 for Company A

BEST COPY AVAILABLE

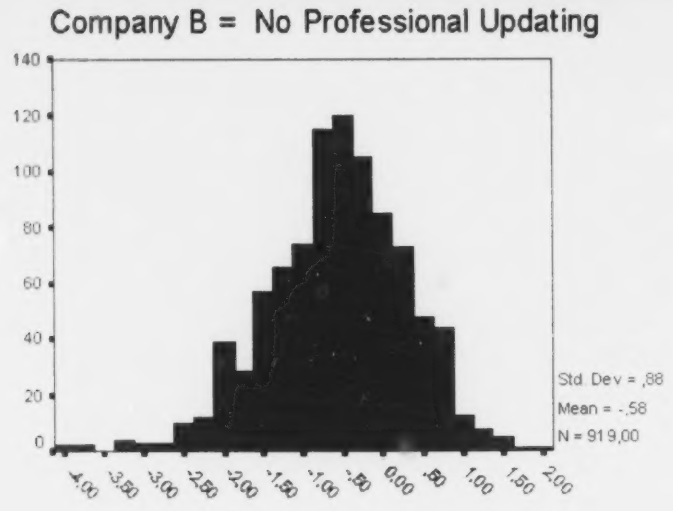


Figure 9

Canonical Discriminant Function 1 for Company B

Figure 10 proposes that new measured Company C needs consulting on different areas of professional drowth. This figure is not trying to tell researcher what exactly needs to be done rather than suggest further investigation. On the other hand, this tools ability to classify Organizational Triggering variables measuring climate needs to be tested more in the near future with time series analysis.

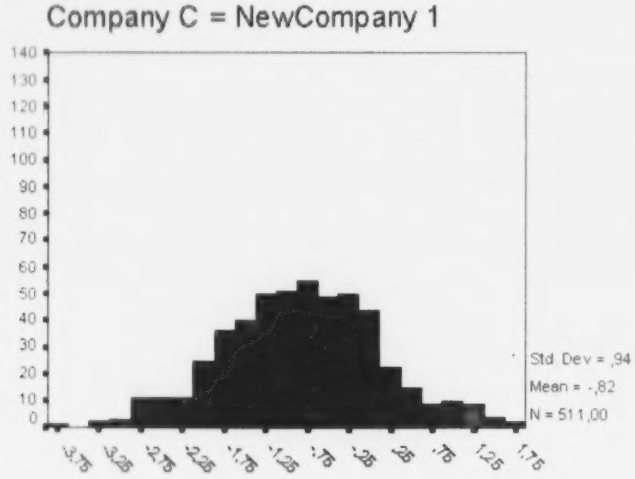


Figure 10

Canonical Discriminant Function 1 for Company C

Conclusions

Main purpose of this study was to show educational researchers how easy and intuitive it is to use non linear Bayesian methods to perform classification on vocational data with BAYDA.

As reader noticed, we used several pages to report linear analysis and only few for non-linear analysis. Of course me must issue that linear analysis was reported at greater accuracy due to fact that non-linear analysis was already presented in the previous chapter (Silander, T. and Tirri, H., 1998). Reader should know that we also spent several hours more to analyze (or produce) linear output than non-linear. With BAYDA one has to make five easy and quick steps to achieve more intelligible result. As stated earlier, BAYDA 1.0 is more like a colleague and tutor, than a tool.

Final conclusion of this study is that linear and non-linear methods support each other depending of the subject of the study. Bayesian approach in the form of BAYDA is still under rapid development. Forthcoming features include opportunity to carry out realtime modelling with different classification scenes. As computing power on desktop computers is likely to increase, one might expect to see in a near future a new arrival of Bayesian applications for educational researchers.

References

Beairisto, B. (1996). Professional Growth and Development: What is it and how we know if it`s working? in Professional Growth and Development. *Career Education Books, Saarijärvi, Finland.*

Honka, J. and Ruohotie, P. (1997). Developing Skills in Organization. *RT Consulting Team, Saarijärvi, Finland.*

Kautto-Koivula, K. (1993). Degree-Oriented Professional Adult Education in the Work Environment. Acta Universitas Tamperensis. *Ser A, vol 390. Tampere.*

Lahti-Kotilainen, L. (1992). Values as Critical Factors in Management Training. *Acta Universitas Tamperensis. Ser A, vol 390. Tampere.*

Rauhala, P. and Ruohotie, P. (1992). Flexible Educational Structure as Development Project for Vocational Education. *University of Tampere, Tampere, Finland.*

Ruohotie, P. (1995). Professional Growth in the Work Environment. *University on Tampere, Tampere, Finland.*

Ruohotie, P. and Grimmet, P. (1994). New Themes for Education in a Changing World. *Career Education Books, Saarijärvi, Finland.*

Ruohotie, P. and Grimmet, P. (1996). Professional Growth and Development. *Career Education Books, Saarijärvi, Finland.*

Silander, T. and Tirri, H. (1998). Bayesian Classification. *Modern Modeling of Professional Growth. RCVE, Hämeenlinna, Finland.*

SPSS Inc. (1997). Advanced Statistics. *SPSS Inc, Chicago, USA.*

Tabachnick, B. and Fidell, L. (1996). Using Multivariate Statistics. *Harper Collins College Publishers, California, USA.*

Tirri, H. (1998). What the Heritage of Thomas Bayes has to offer for Modern Educational Research? *Modern Modeling of Professional Growth. RCVE, Hämeenlinna, Finland.*